Robust Training and Predictive Uncertainty in Deep Learning

Speaker: Sunil Thulasidasan

sunil@lanl.gov

Los Alamos NATIONAL LABORATORY EST. 1943

Workshop on AI and Tensor Factorization for Physical, Chemical, and Biological Systems, Santa Fe, 2019

UNIVERSITY OF WASHINGTON ELECTRICAL ENGINEERING

sunil@lanl.gov

Outline

- Supervised deep learning in the presence of label noise
- Improving predictive uncertainty of deep models

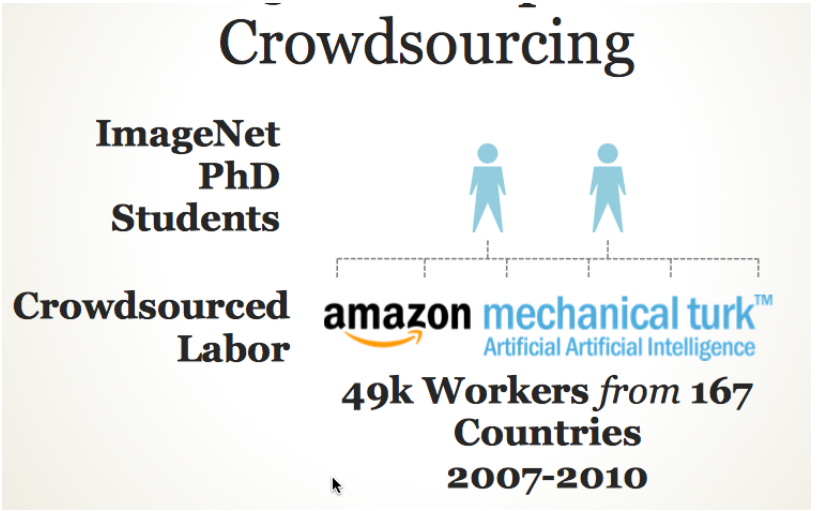
sunil@lanl.gov

A Practical Challenge for Deep Learning

State-of-the-art models require *large amounts of* ***clean****, annotated data.*

sunil@lanl.gov

Annotation is labor intensive!



Slide from Fei-Fei Li and Jia Deng

ImageNet: 15 million labeled images; over 20,000 classes

The data that transformed AI research—and possibly the world (D. Gershgorn, quartz, magazine, 2017)

- 49k workers
- 167 countries
- 2.5 years to complete!

sunil@lanl.gov

Approaches to large-scale labeling

- Crowdsource at scale – labor intensive, but relatively cheap

amazon beta mechanical turk

- Use weak labels from queries, user tags and pre-trained classifiers

sunil@lanl.gov

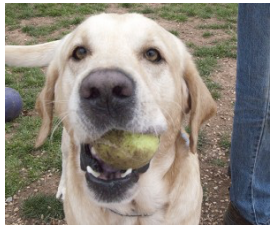
Approaches to large-scale labeling

- Crowdsource at scale – labor intensive, but
- Use weak labels from queries, user tags and pre-trained classifiers

Both approaches can lead to significant labeling errors!

Slide credit: S Guo et al '2018

- Label noise is an inconsistent mapping from features ***X*** to labels ***Y***



Dog ✓

Dog ✓

Dog 🤔

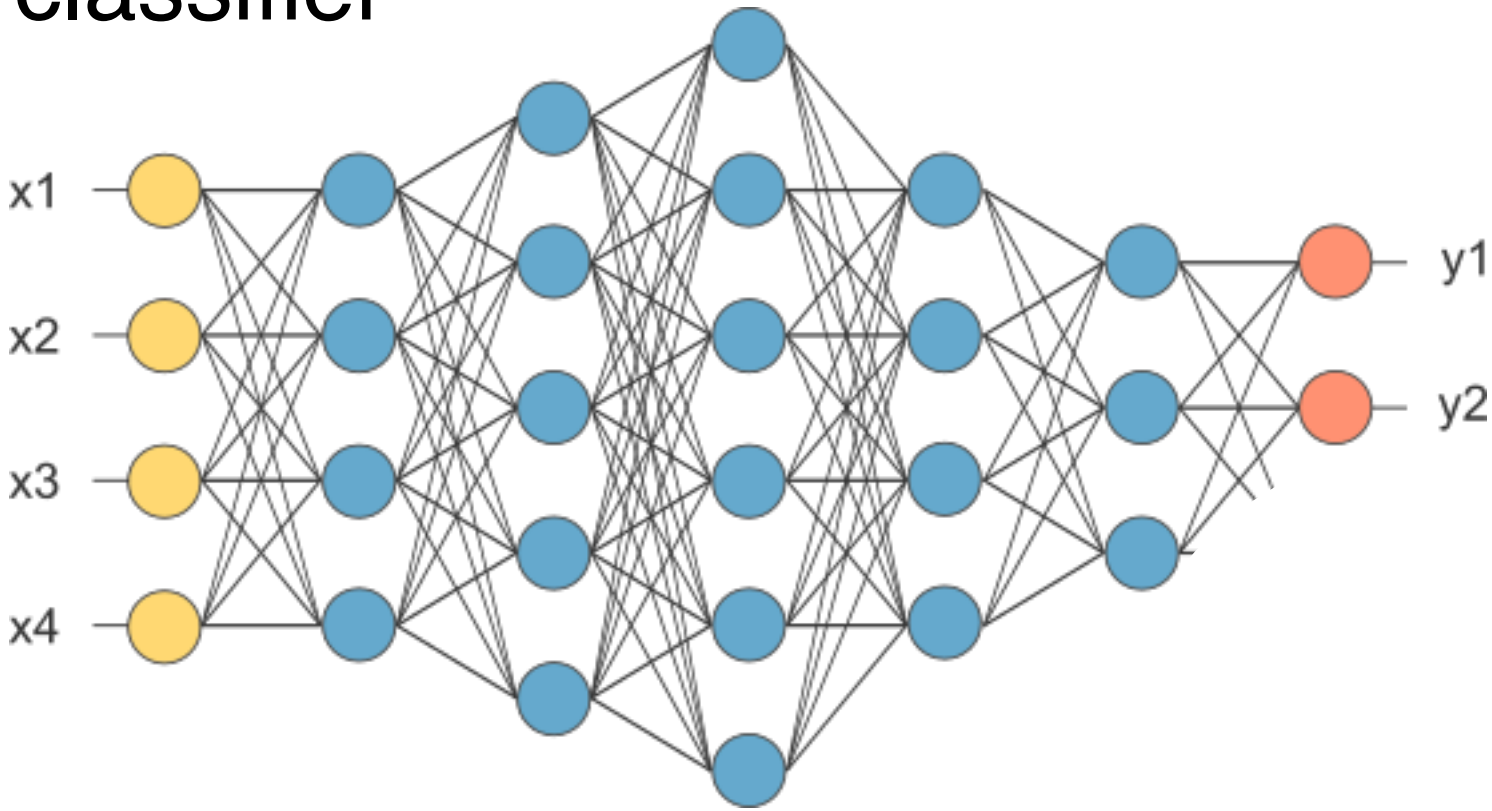
The Deep Abstaining Classifier (DAC)

Approach: Use learning difficulty on incorrectly labeled or confusing samples to defer on learning -- ***"abstain"*** -- till correct mapping is learned.

sunil@lanl.gov

DAC Overview

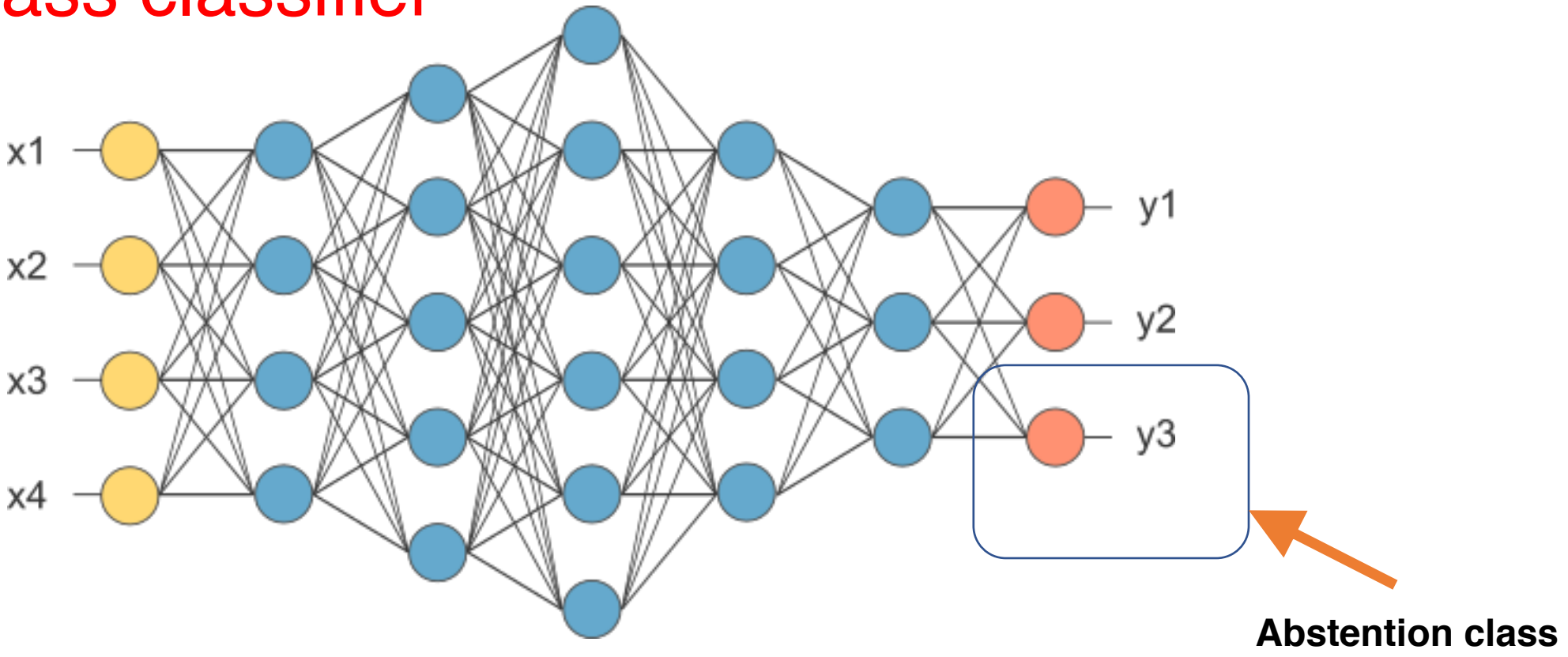
2 class classifier



sunil@lanl.gov

DAC Overview

2 + 1 class classifier



sunil@lanl.gov

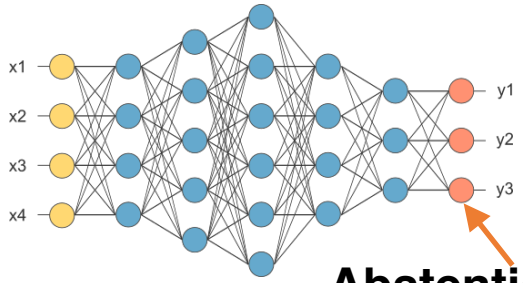
Training a Deep Abstaining Classifier

$$\mathcal{L}(x) = (1 - p(x)_{k+1}) \left(-\sum_{i=1}^{k} t(x)_i \log \frac{p(x)_i}{1 - p(x)_{k+1}} \right) + \alpha \log \frac{1}{1 - p(x)_{k+1}}$$

Cross entropy as usual

sunil@lanl.gov

Training a Deep Abstaining Classifier



Abstention class

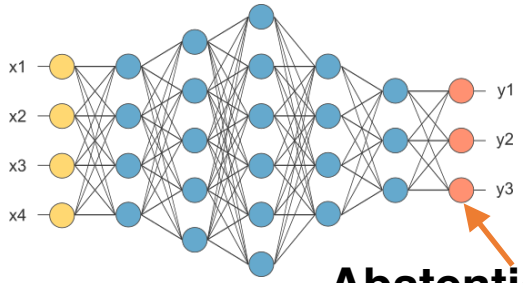
$$\mathcal{L}(x) = (1 - p(x)_{k+1}) \left(-\sum_{i=1}^{k} t(x)_i \log \frac{p(x)_i}{1 - p(x)_{k+1}} \right) + \alpha \log \frac{1}{1 - p(x)_{k+1}}$$

Encourages abstention

Cross entropy over actual classes

sunil@lanl.gov

Training a Deep Abstaining Classifier



Abstention class

$$\mathcal{L}(x) = (1 - p(x)_{k+1}) \left(-\sum_{i=1}^{k} t(x)_i \log \frac{p(x)_i}{1 - p(x)_{k+1}} \right) + \alpha \log \frac{1}{1 - p(x)_{k+1}}$$

Encourages abstention

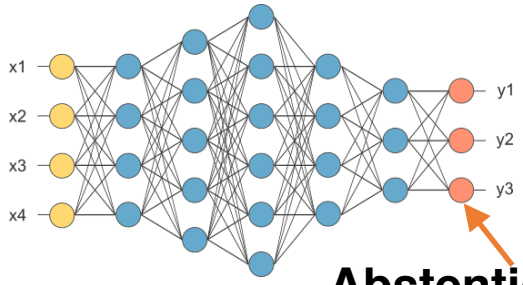
Cross entropy over actual classes

Automatically tuned during learning.

Penalizes abstention

sunil@lanl.gov

Training a Deep Abstaining Classifier



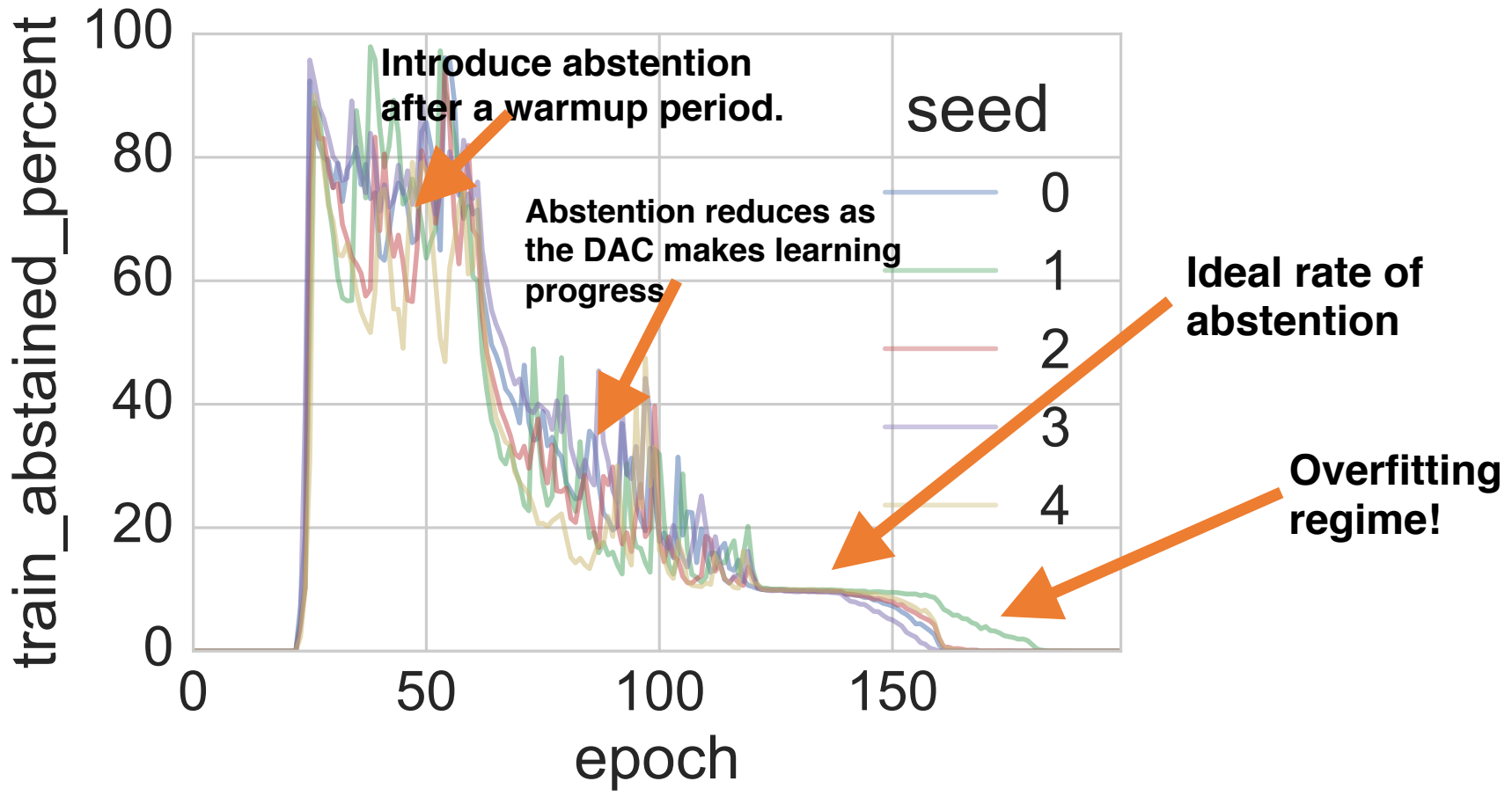
$$\mathcal{L}(x)_{\mathrm{DAC}} = (1 - p_{k+1}) \left(- \sum_{i=1}^{k} t_i \log \frac{p_i}{1 - p_{k+1}} \right) + \alpha \log \frac{1}{1 - p_{k+1}}$$

Absence of the extra abstention class exactly recovers the standard loss.

$$p_{k+1} = 0 \Rightarrow \mathcal{L}_{\mathrm{DAC}} = \left(- \sum_{i=1}^{k} t_i \log p_i \right) = \mathcal{L}_{standard}$$

sunil@lanl.gov

Abstention Dynamics



Abstained percent on training set vs epoch with **10% label noise**.

sunil@lanl.gov

Training Protocol

Training protocol:

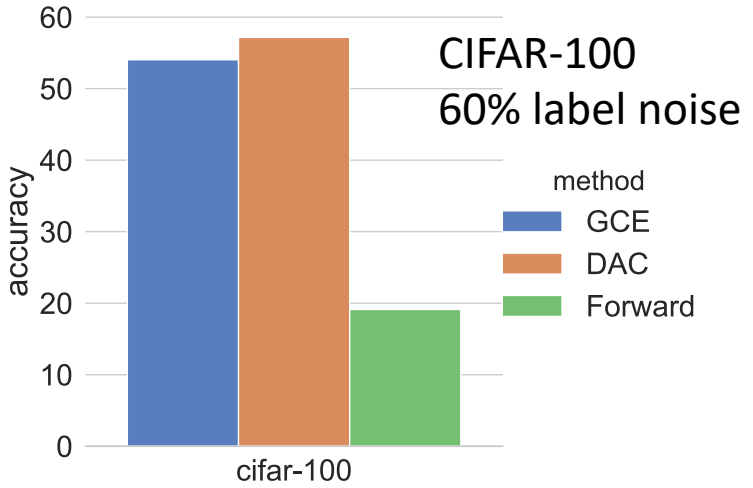
- Use DAC to identify label noise.
- Eliminate noisy data
- Retrain on cleaner set.

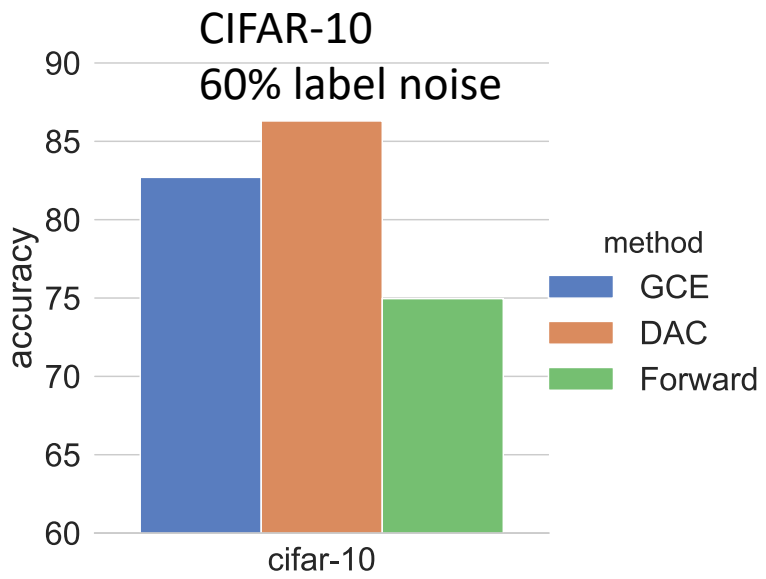
sunil@lanl.gov

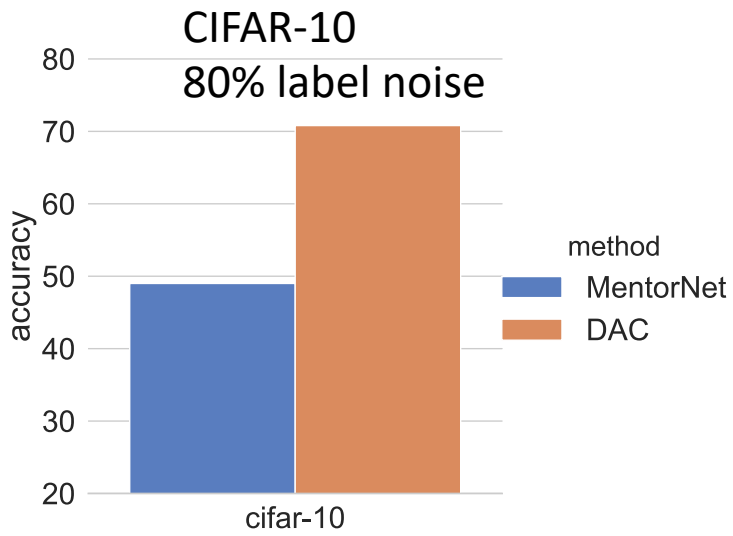
The DAC gives state-of-art results in experiments with arbitrary label noise

Training protocol:

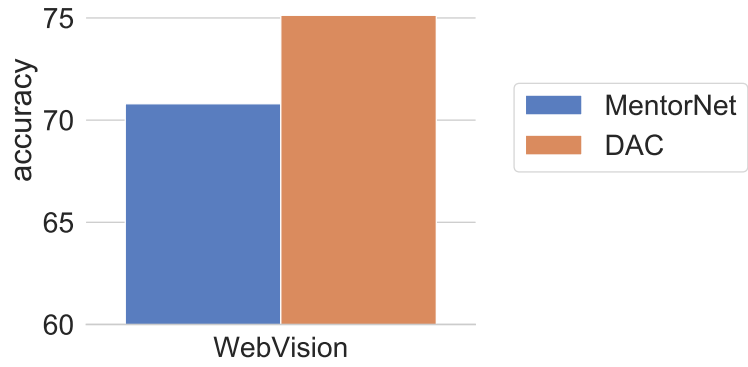
- **Use DAC to identify and eliminate label noise.**
- **Retrain on cleaner set.**







WebVision: Real-world noisy dataset. ~2.4M images. ~35-40% label noise



GCE: Generalized Cross-Entropy Loss (Zhang et al NIPS '18); Forward (Patrini et al, CVPR '17); MentorNet (Li et al, ICML '18)

Abstention in the Presence of Systematic Label Noise

sunil@lanl.gov

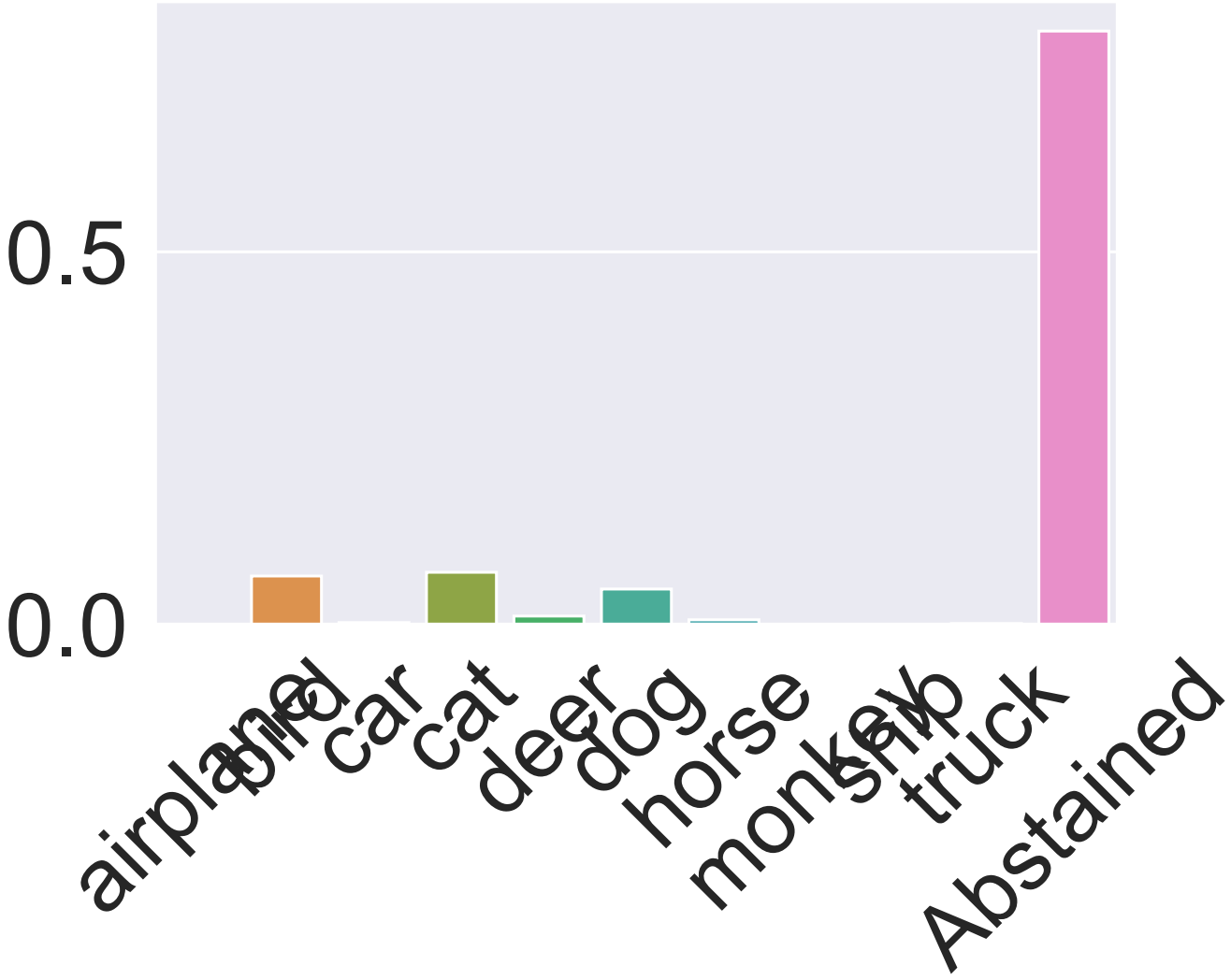
Abstention in the presence of Systematic Label Noise: The Random Monkeys Experiment

All the monkey labels in the training set (STL-10) are randomized.

Can the DAC learn that images containing monkey features have unreliable labels and abstain on monkeys in the ***test set?***

Random Monkeys: DAC Predictions on Monkey Images

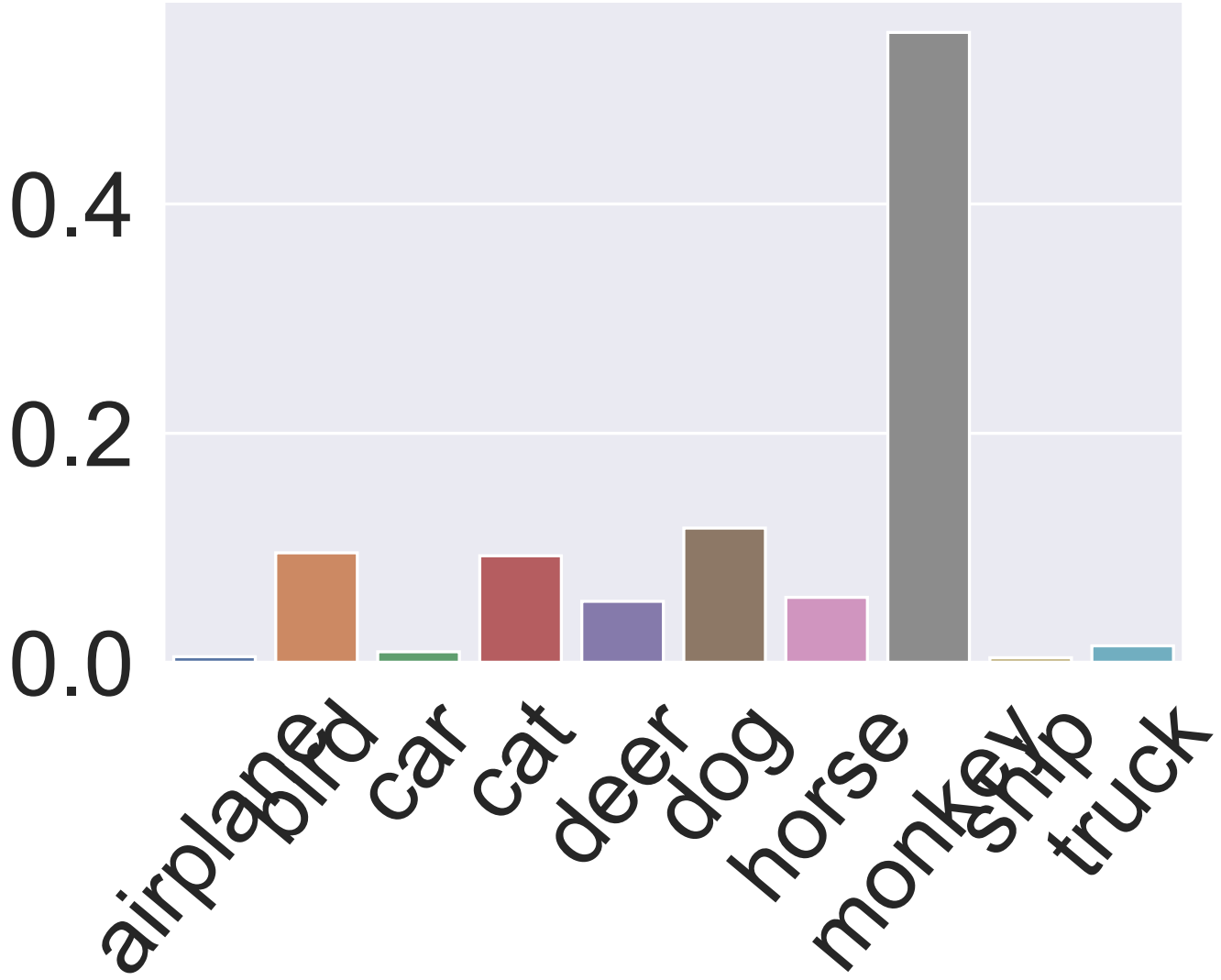
sunil@lanl.gov



The DAC abstains on most of the monkeys in the test set!

sunil@lanl.gov

Random Monkeys: Class Distribution of Abstained Images



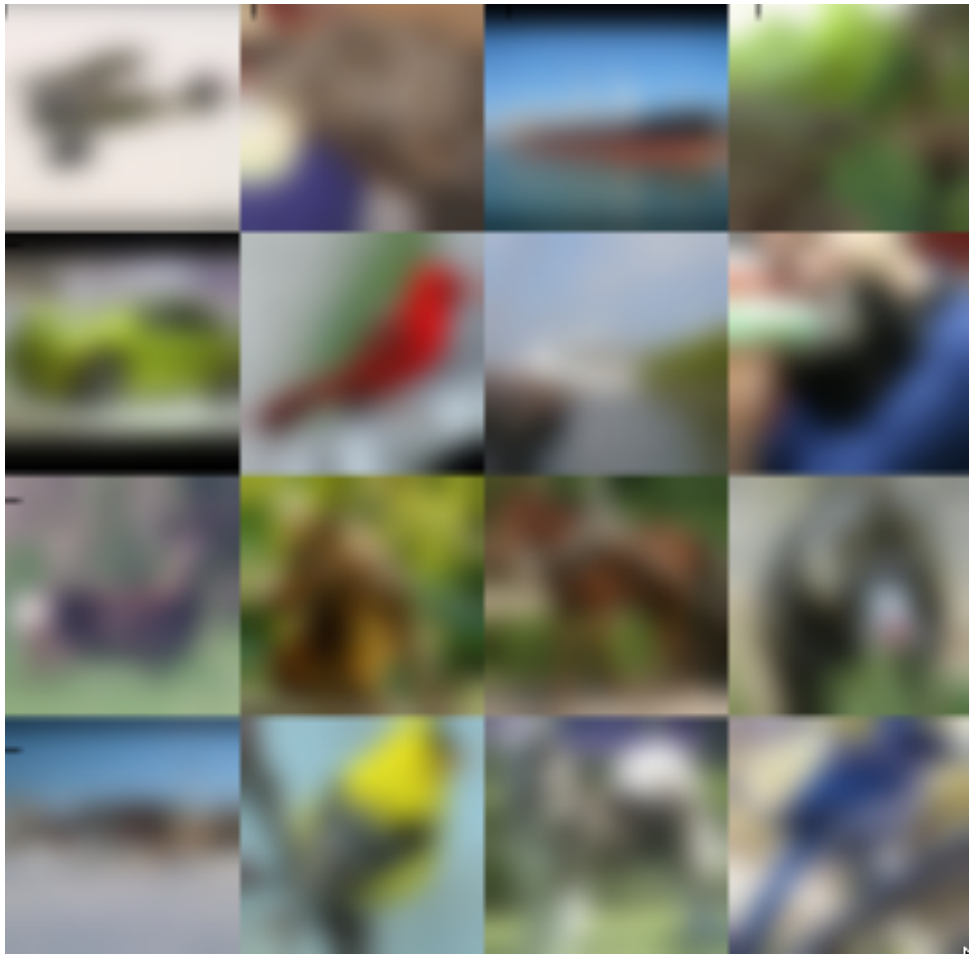
Most of the abstained images are monkeys.

sunil@lanl.gov

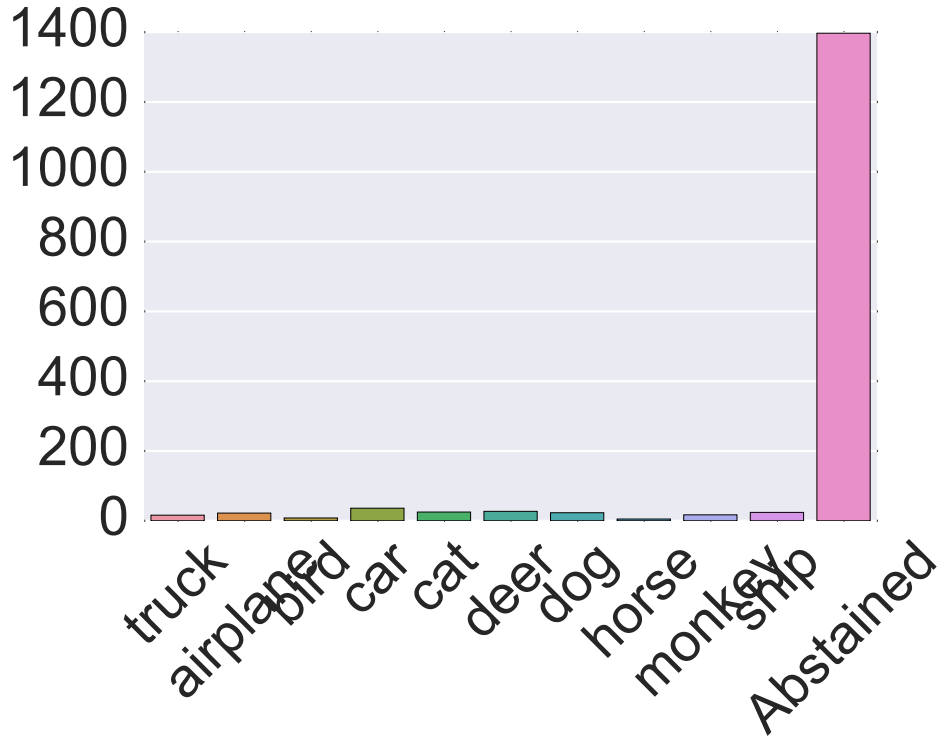
Image Blurring

Blur a subset (20%) of the images in the training set and randomize labels

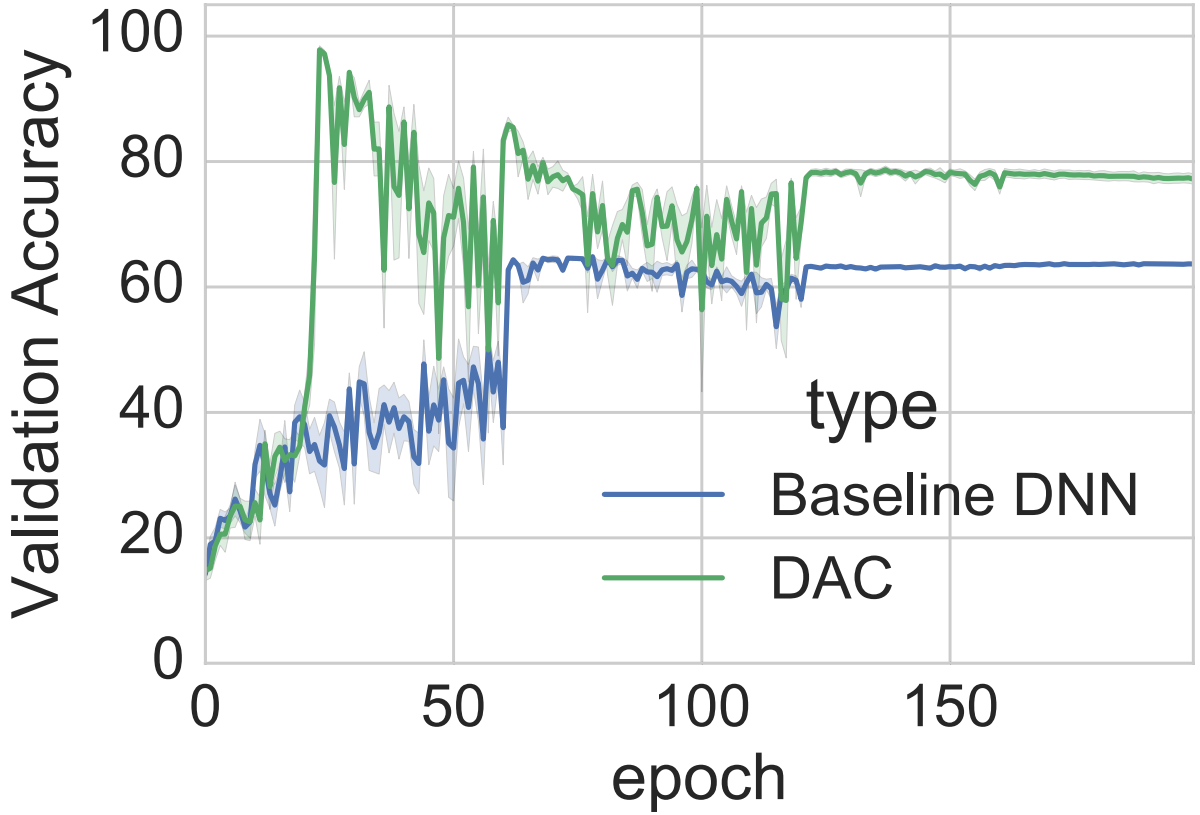
Will the DAC learn to abstain on blurred images in the test set?



DAC Behavior on Blurred Images



DAC abstains on most of the blurred images in the test set



For DAC, validation accuracy is calculated on non-abstained samples.

sunil@lanl.gov

DAC can correctly learn features associated with label noise

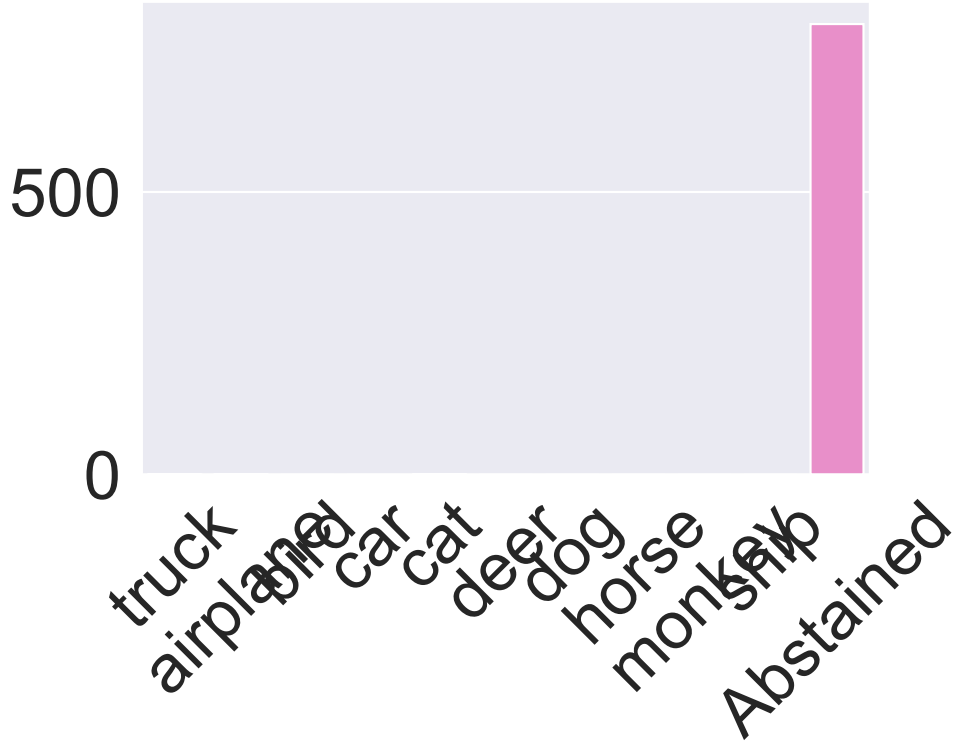
A smudge is added to 10% of the training and test set.

Labels of smudged images in the training set are randomized.

sunil@lanl.gov

DAC can correctly learn features associated with label noise

As a simple test of feature learning for label noise, we add a *smudge* to 10% of the training set and randomize the labels on the smudged images



Predictions on smudged Images (test set)

sunil@lanl.gov

What does the DAC “see” while abstaining?

An abstained image in the test set. The smudge completely dominates the feature saliency map.

Same image without the smudge. Class features are more salient and the class is correctly predicted.

Convolutional filter visualization using guided grad-cam (Selvaraju et al CVPR ‘17)

sunil@lanl.gov

What does the DAC “see” while abstaining?

Convolutional filter visualization using class-activation maps for an abstained monkey image

The DAC abstains based on monkey features.

In other words, it has formed a mapping from monkey features to abstention class due to label noise!

sunil@lanl.gov

What does the DAC “see” while abstaining?

Abstained monkey image

What the DAC “saw”

sunil@lanl.gov

Conclusions

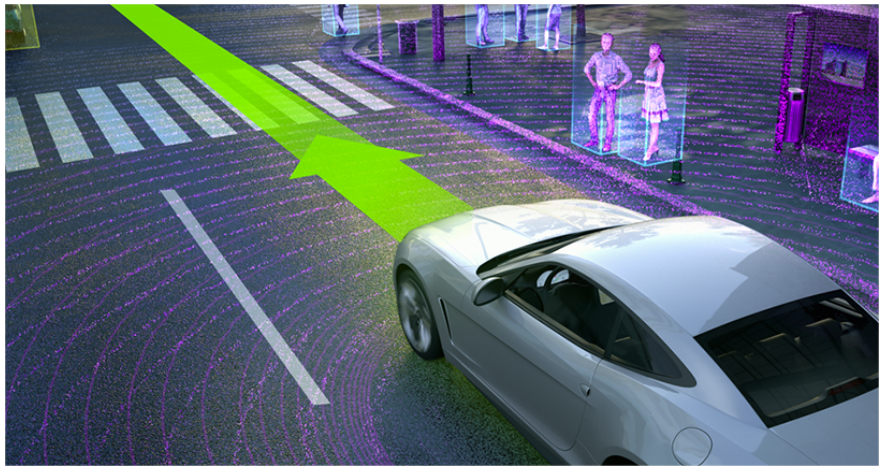
Code available at https://github.com/thulas/dac-label-noise

- **Abstention training is an effective way to clean label noise in a deep learning pipeline**.
- Abstention can also be used as a ***representation learner*** for label noise.
  - Especially useful for interpretability in "don't-know" decision situations.
- **Publication:** S. Thulasidasan, T. Bhattacharya, J. Bilmes, G. Chennupati, J. Yusof: ***"Combating Label Noise in Deep Learning using Abstention" ,*** ICML 2019.

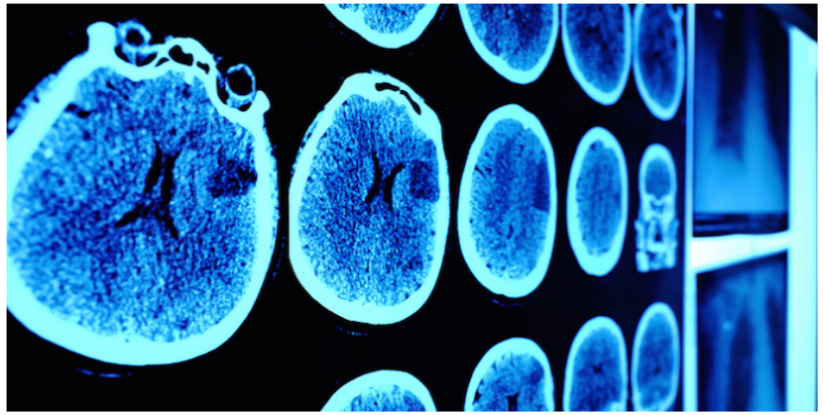
Improving Predictive Uncertainty in Deep Learning

sunil@lanl.gov

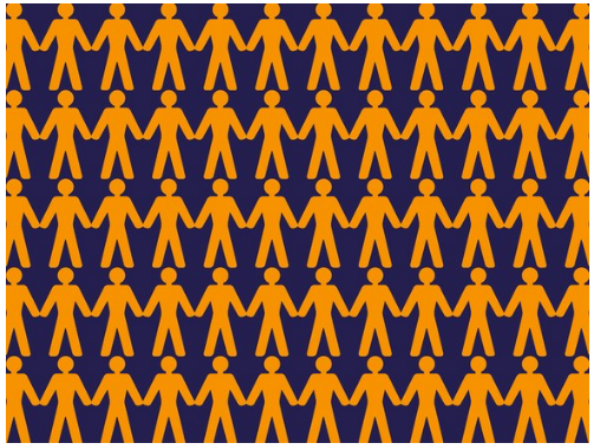
Can we trust deep models to make high risk decisions?



A deep learning tool developed by Google accurately identified breast cancer in pathology slides and reduced average slide review time.



COURTS ARE USING AI TO SENTENCE CRIMINALS. THAT MUST STOP NOW



sunil@lanl.gov

Uncertainty and Over-confidence in Deep Models

- Traditional way to model uncertainty: use the score associated with a model and possibly a threshold

- Problem: Deep models can very confidently give the wrong predictions!

sunil@lanl.gov

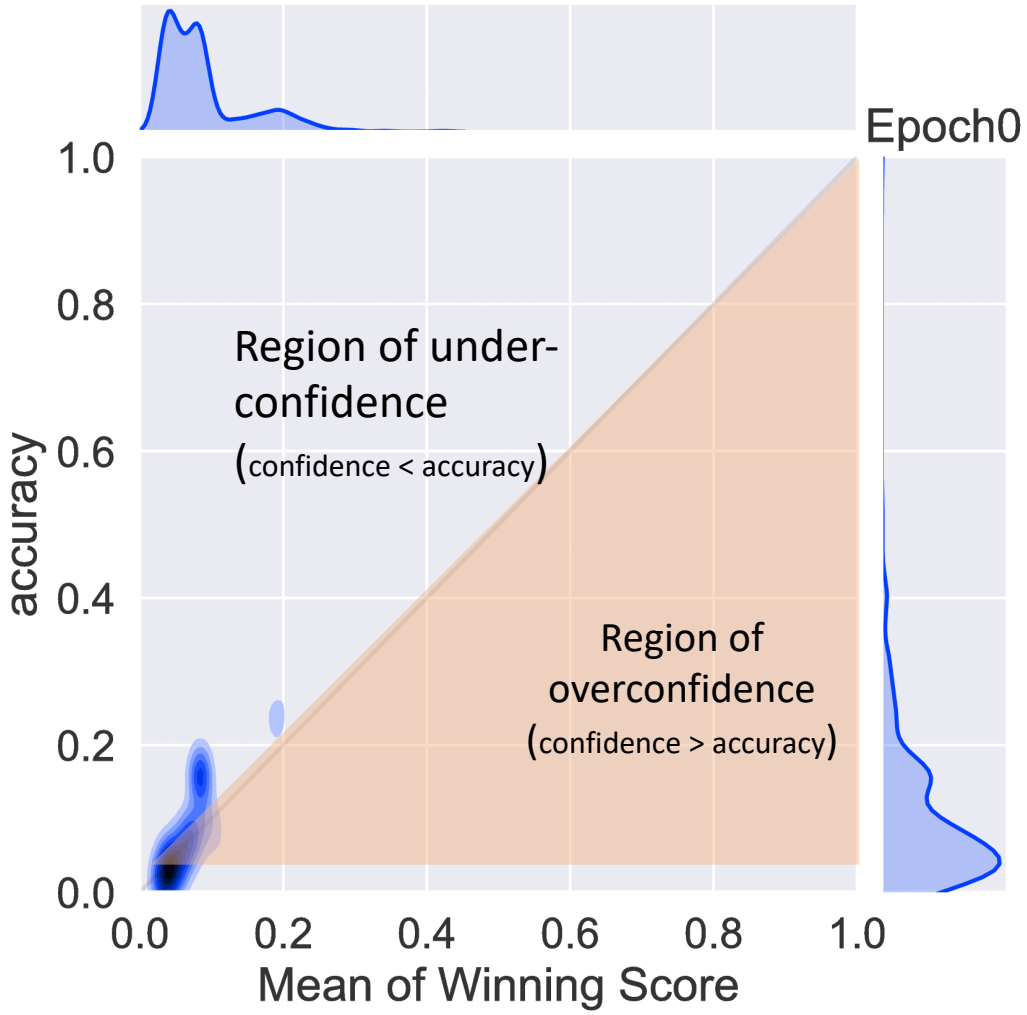
Modern Deep Network are mis-calibrated

- In a well calibrated classifier, predicted score should reflect the probability of being correct.

- Modern deep neural networks tend to be mis-calibrated

sunil@lanl.gov

Typical Training of DNNs leads to miscalibration and overconfidence



VGG—16

CIFAR-100

[animation link]

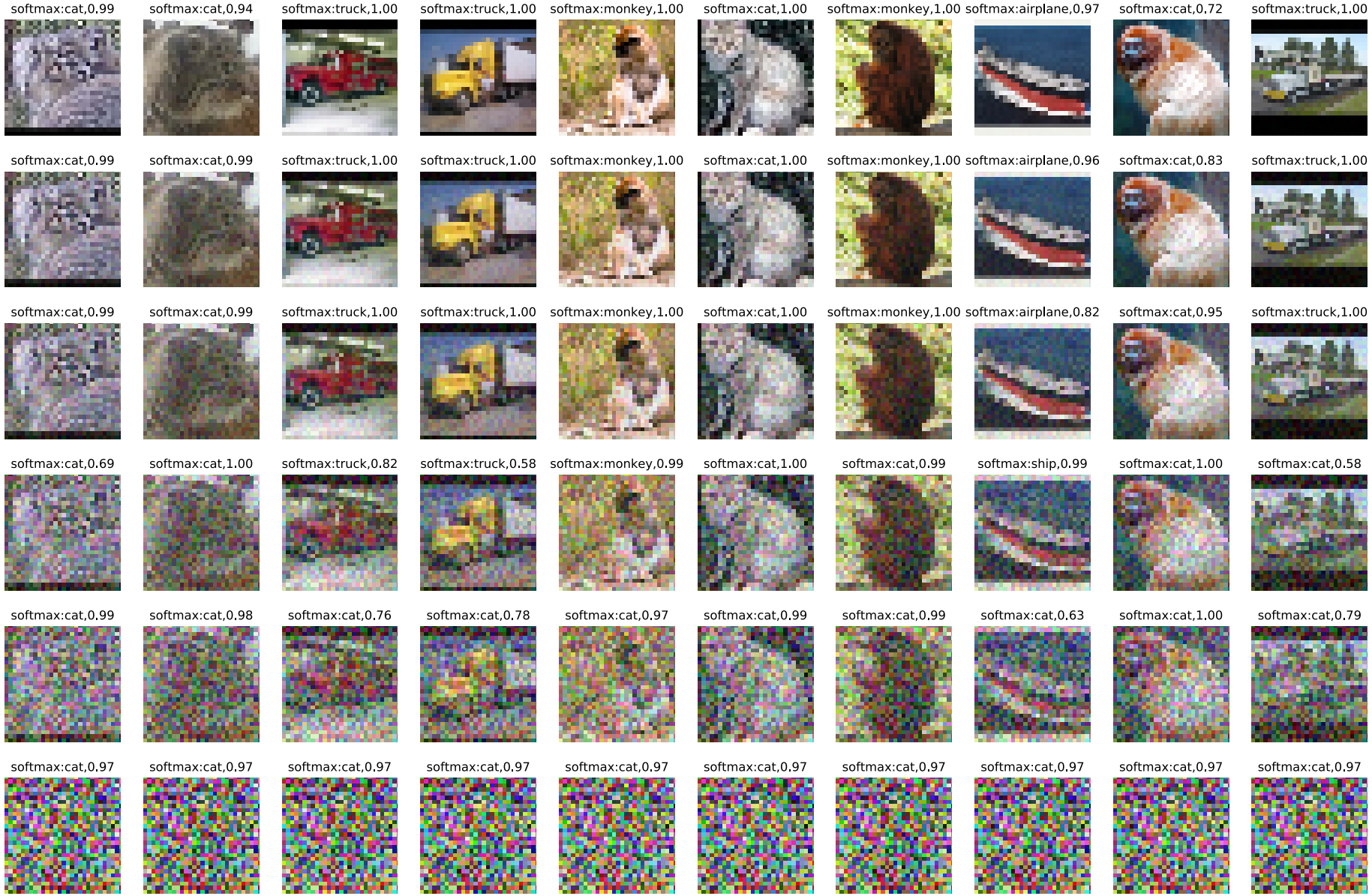
DNNs can predict confidently on random noise!

A DNN image classifier predicts this as ‘cat’ with 99% confidence!

Image Perturbation Experiment

Given an input image $\mathbf{X} \in \mathbb{R}^m$, we choose a random vector $\mathbf{d} \in \mathbb{R}^m$ (where $d_i \sim U(0,1)$), and perturb $\mathbf{X}$ as follows: $\mathbf{X}' = \mathbf{X} + \alpha\hat{\mathbf{d}}$.

Predictions on perturbed images



sunil@lanl.gov

The Rectified Linear Unit (ReLU): A source of overconfidence in deep models

- Rectified linear units are the most widely used hidden layer non-linearities.

- Recent work (Hein et al 2018) shows that ReLU models can be pathologically overconfident on points far away from training manifold.

sunil@lanl.gov

ReLU: A source of overconfidence in deep models?

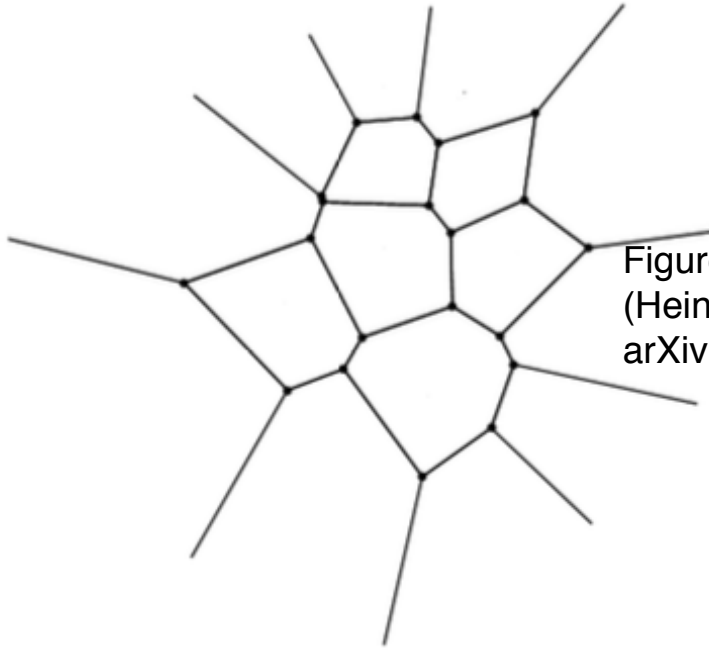


Figure from (Hein et al arXiv '18)

Figure 1: A decomposition of $\mathbb{R}^2$ into a finite set of polytopes. The outer polytopes extend to infinity. This is where ReLU networks realize arbitrarily high confidence predictions.

sunil@lanl.gov

Another Source of Over-confidence: Training with hard labels

- Hard labels have all the probability mass in one class
- Thus the DNNs, are in some sense, *trained to become overconfident*.

sunil@lanl.gov

Another Source of Over-confidence: Training with hard labels

- Hard labels have all the probability mass in one class
- Thus the DNNs, are in some sense, *trained to become overconfident*.
- Would soft labels temper overconfidence?
  - How does one generate soft-labels in a principled manner?

sunil@lanl.gov

A closer look at recent data augmentation techniques

- Mixup is a data augmentation technique for images where samples *and their labels* are convexly combined during training (Zhang et al ICLR 2018); shown to improve classification performance.

Mixup Training

0.108

0.892

Convexly combine images and labels

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j$$

New Label:

Dog	Ship	Other…
0.108	**0.892**	**0….**

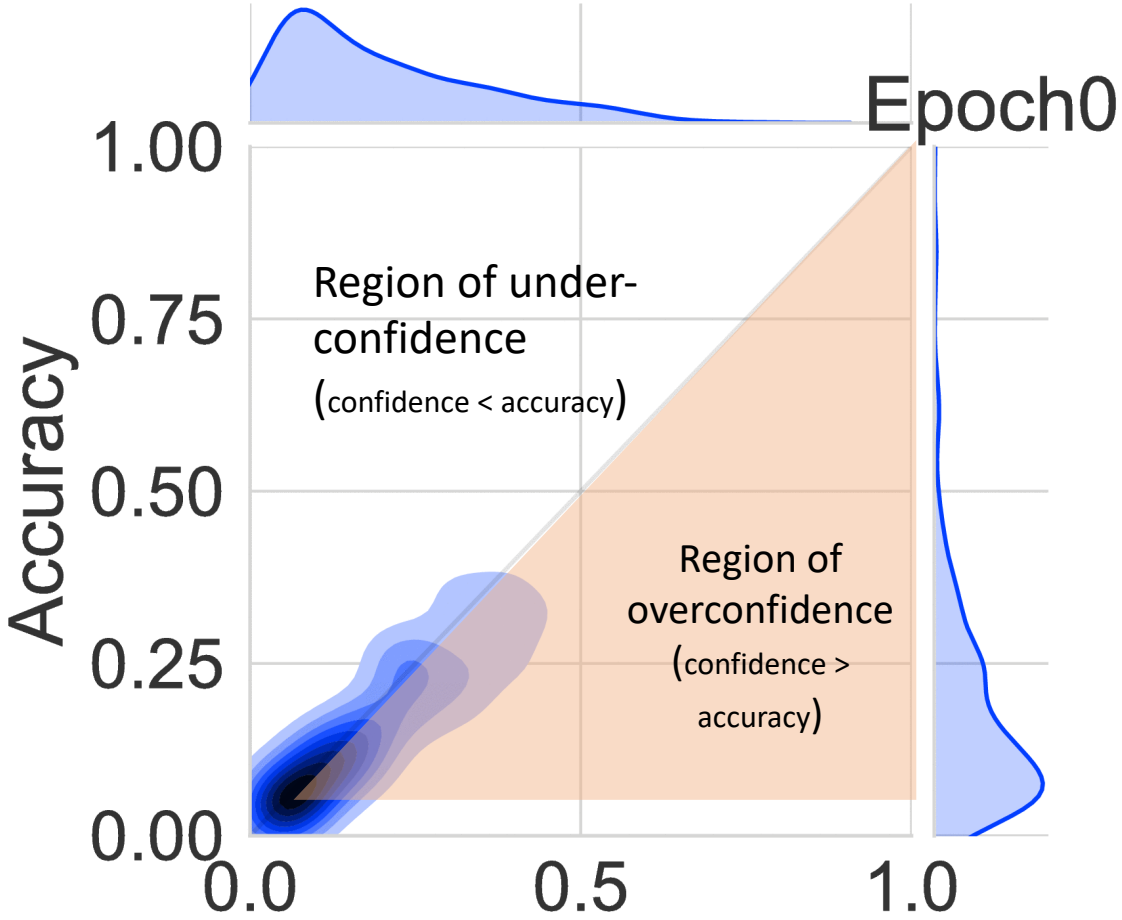
sunil@lanl.gov

A closer look at recent data augmentation techniques

- Our hypothesis was training on hard labels (zero-entropy distributions) leads to overconfidence.
- If above is true, techniques like mixup – where the training labels are smoothened out due to convex mixing – should temper overconfidence

sunil@lanl.gov

Mix-up Training Leads to well-calibrated models

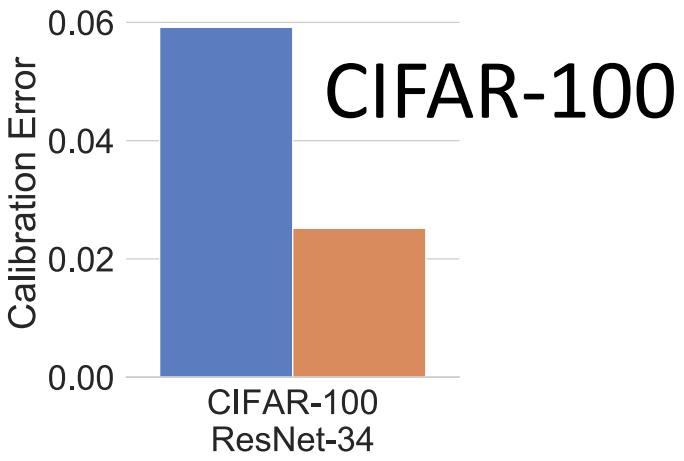


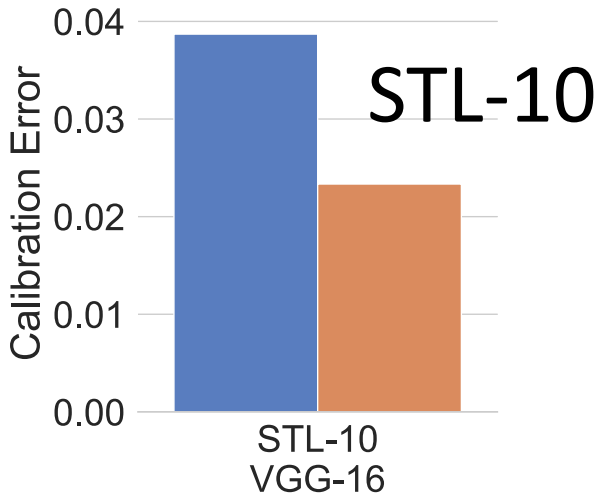
VGG—16

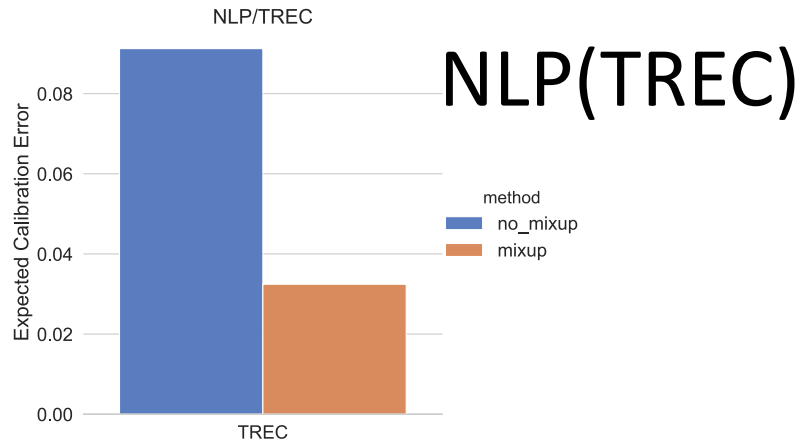
CIFAR-100

animation

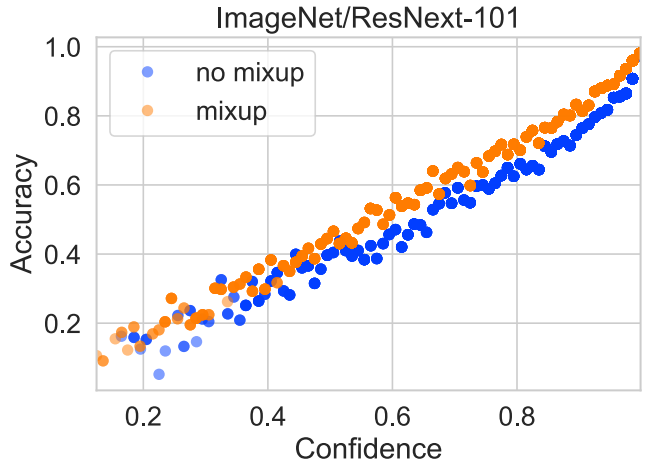
Mixup training reduces calibration error

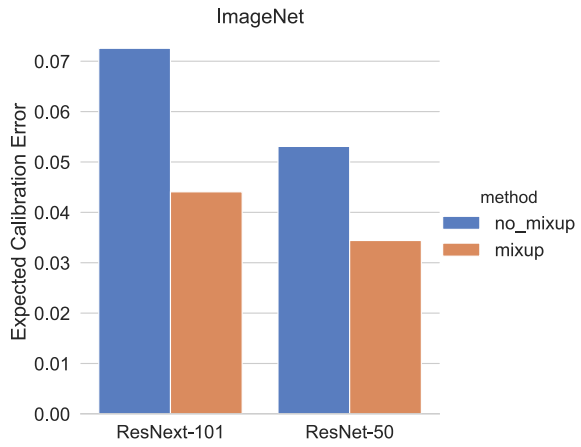


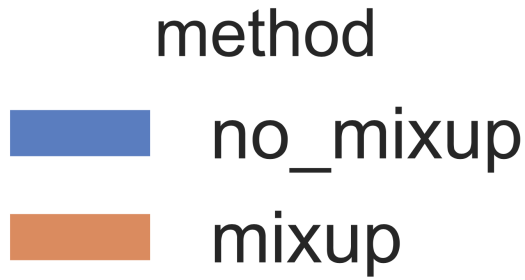




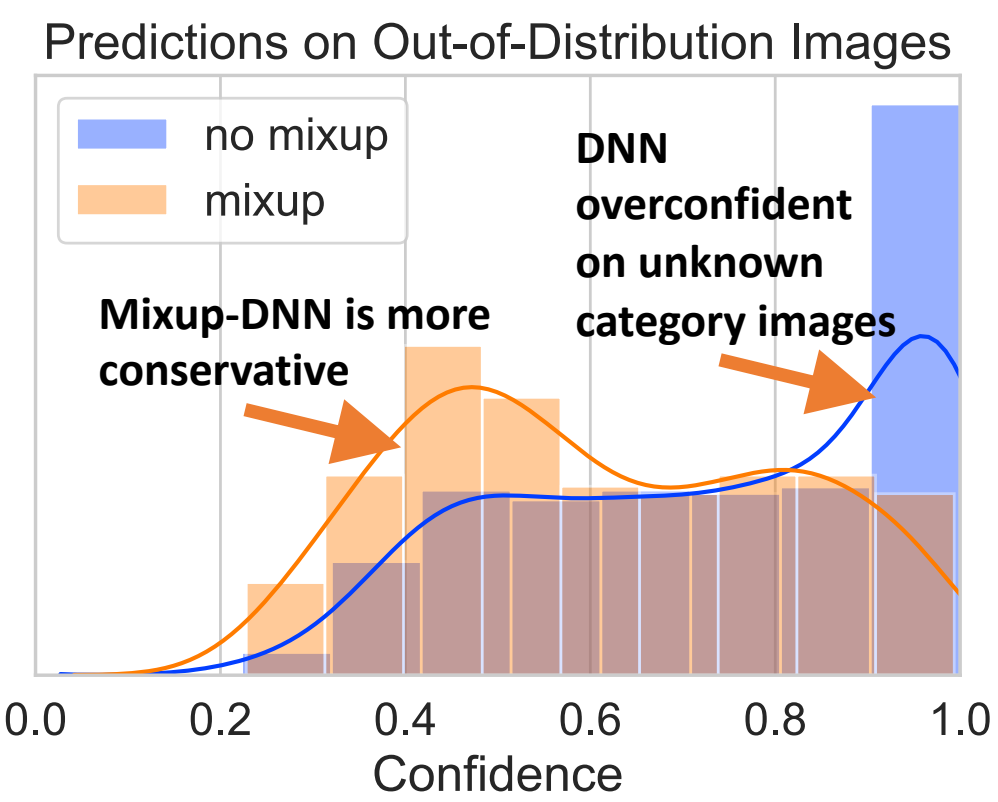
ImageNet

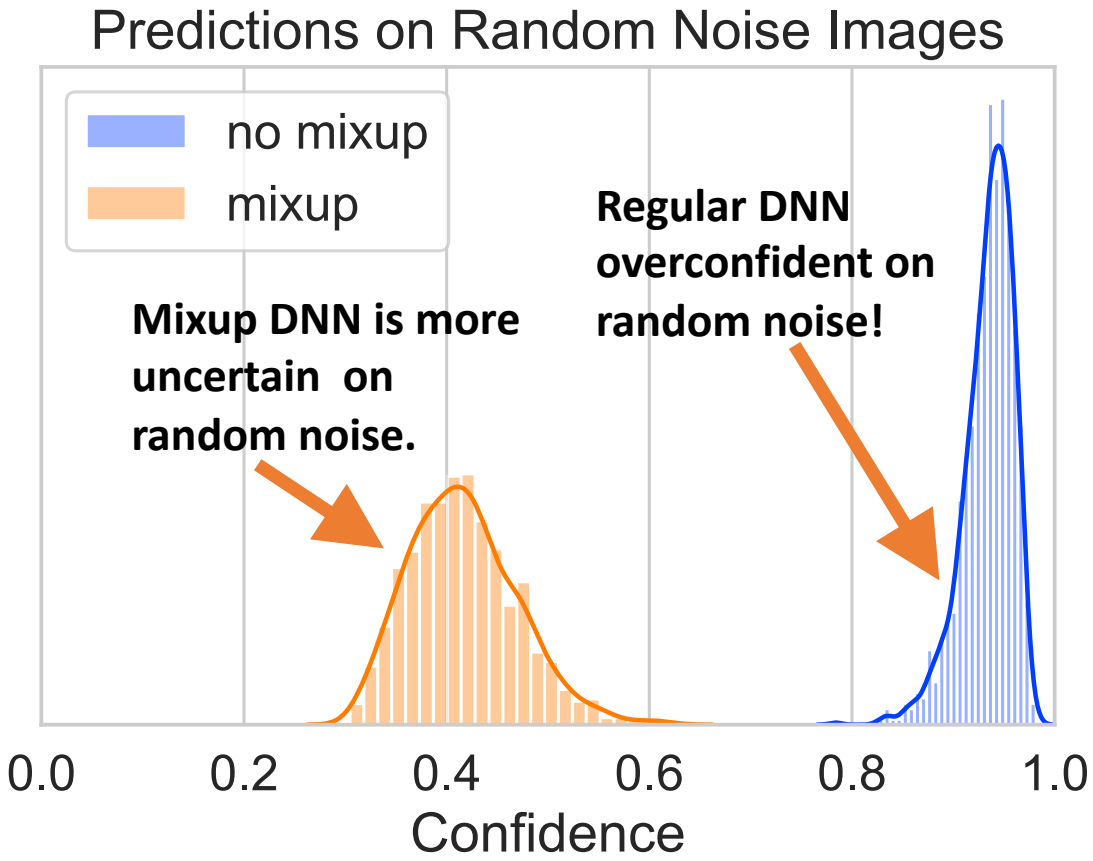






Mixup training shows improved uncertainty on open-set and random-noise images





Distribution of winning scores

sunil@lanl.gov

Conclusions

- **Label smoothing in the inputs are thus an effective and *efficient* way to improve predictive uncertainty in deep neural networks**

- **Publication:** S. Thulasidasan, , G. Chennupati, J. Bilmes, T. Bhattacharya, S. Michalak : **"On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks,** To appear in NeurIPS 2019.